Topical organization of user comments and application to content recommendation
Vidit Jain, Esther Galbrun

To cite this version:
Vidit Jain, Esther Galbrun. Topical organization of user comments and application to content recommendation. International Conference on World Wide Web, May 2013, Rio de Janeiro, Brazil. 10.1145/2487788.2487812 . hal-01399275

HAL Id: hal-01399275
https://hal.science/hal-01399275
Submitted on 25 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Topical Organization of User Comments and Application to Content Recommendation

Vidit Jain
Yahoo! Labs Bangalore, India
viditj@yahoo-inc.com

Esther Galbrun
Department of Computer Science and Helsinki Institute for Information Technology HIIT, University of Helsinki, Finland
galbrun@cs.helsinki.fi

ABSTRACT
On a news website, an article may receive thousands of comments from its readers on a variety of topics. The usual display of these comments in a ranked list, e.g., by popularity, does not allow the user to follow discussions on a particular topic. Organizing them by semantic topics enables the user not only to selectively browse discussions on a topic, but also to discover other significant topics of discussion in comments. This topical organization further allows to explicitly capture the immediate interests of the user even when she is not logged in. Here we use this information to recommend content that is relevant in the context of the comments being read by the user. We present an algorithm for building such a topical organization in a practical setting and study different recommendation schemes. In a pilot study, we observe these comments-to-article recommendations to be preferred over the standard article-to-article recommendations.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous

Keywords
Algorithm; User Generated Content; Recommendation;

1. INTRODUCTION
Consider the example of an article about a recent tornado in Alabama. The main topics appearing in the comments for this article are displayed in a compact and interactive interface (see Figure 1(a)). Each block corresponds to one cluster, summarized by a tag cloud of the entities that occur in it. In this illustration, the user is reading the comments that discuss the emergency responses of the Federal Emergency Management Agency (FEMA) after this natural disaster (cluster highlighted in a darker shade). At a different time, this user may be reading other comments that question the legitimacy of government’s spending of billions of dollars as foreign aid. In the first case, the user is likely to be more interested in reading about the responses after similar natural disasters, such as hurricane Katrina, as opposed to information on foreign aid efforts. Existing systems either use only the article text along with a global personalization model to recommend articles. These systems fail to distinguish between the above two different information needs of a single user. Our system addresses this issue by recommending an article relevant to the specific context of a semantically coherent cluster of comments (Figure 1(b)).

Our conclusions do not compete with, but complement, other approaches that model user profiles and user similarities for providing personalized recommendations.

2. TOPICAL ORGANIZATION
Building topical organization for user comments has its unique challenges. The difference in the language skills and seriousness of comment writers leads to a huge variation in the language and content across comments. The standard natural language processing and information extraction systems perform poorly on these noisy, short pieces of text. Also, for popular articles, the comments arrive at a high rate and the topics of discussions evolve rapidly over time. On a typical news website, the desired system need to compute these topics for thousands of articles every few minutes. Another constraint is due to the actual representation of the extracted topics to the users – in order to provide a consistent user experience across different articles, the hosting website prescribed that at most five different clusters...
Table 1: Comparison of clustering algorithms. Average values of the entropy of intra-cluster term frequencies (TE), per article inter-cluster overlap between top-5 terms (TO), entropy of topic sizes (CE), and fraction of comments assigned a topic (COV).

<table>
<thead>
<tr>
<th></th>
<th>TE</th>
<th>TO</th>
<th>CE</th>
<th>COV</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means</td>
<td>0.542</td>
<td>4.353</td>
<td>0.120</td>
<td>0.474</td>
</tr>
<tr>
<td>METIS</td>
<td>0.517</td>
<td>23.003</td>
<td>0.858</td>
<td>0.511</td>
</tr>
<tr>
<td>EBC</td>
<td>0.511</td>
<td>14.960</td>
<td>1.912</td>
<td>0.505</td>
</tr>
</tbody>
</table>

Table 3: Observations from the pilot study. P-values computed for 1-sided χ² statistical significance test.

<table>
<thead>
<tr>
<th>Recommendation schemes</th>
<th>Accepted hypothesis</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA &gt; AA</td>
<td></td>
<td>0.003</td>
</tr>
<tr>
<td>CA &gt; CCA</td>
<td></td>
<td>0.031</td>
</tr>
</tbody>
</table>

Figure 3: Observations from the pilot study. P-values computed for 1-sided χ² statistical significance test.

3. CONTENT RECOMMENDATION

We consider three recommendation schemes (see Figure 2). The traditional setting for article recommendation (AA) selects the best matching article for a given source article. The second scheme (CA) selects the best matching article for the comments belonging to a given source cluster of comments. Finally, the third scheme (CCA) finds the best matching cluster of comments and selects the parent article of this matching cluster as the target recommendation. These schemes are implemented using a retrieval engine similar to Lucene, which allows a rapid indexing of large volumes of textual data. We use the standard tf-idf representation for the news articles. However, since the comments in the topical clusters show unusual repetitions of terms, we use a slightly different tf term for clusters of comments, i.e., \( \text{tf}_{<d,c}(t, c) = \frac{\# \text{ comments in cluster } c \text{ containing term } t}{\# \text{ comments in cluster } c} \).

The inverse document frequencies of different terms are computed separately for articles and clusters, and denoted by \( \text{idf}_{<d}(t) \) and \( \text{idf}_{<c}(t) \). The relevance of an indexed document \( d \) to a source document \( s \) is then defined as \( R(s, d) = \sum_{t \in d} \text{tf}_{<d}(t, s) \times \text{tf}_{RS}(t, d) \times \text{idf}_{<d}(t)^2 \), where \( t \in \{ art, cld \} \) is the type of the source document and \( \text{tf}_{RS} \) the term frequency computed by the retrieval engine.

We collected the comments for 1884 popular articles (each with at least 500 user comments) from Yahoo! News. In all, we obtained about 9000 clusters (over 2 million comments), for each of which we generated the recommendations using the above schemes. In our pilot study, we first asked users to read comments from a (source) topic. Then they were shown recommendations from each of the above schemes, and asked to choose the most relevant one in the context of the source cluster. As seen in Figure 2 CA is preferred significantly more than both of the AA and CCA schemes. We observed that using topical clusters as sources enables us to identify engaging topics that are latent in the article but are made explicit in the comments. We further investigated machine-learned ensemble techniques that select one of these three recommendations based on the source document. As shown in Table 2 a Gaussian-process based ensemble method was found to perform the best.

Table 2: Improvement (over CA) in predicting user preferences using ensemble methods.

<table>
<thead>
<tr>
<th>Ensemble Method</th>
<th>Improvement (over CA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-class SVM</td>
<td>-3.6%</td>
</tr>
<tr>
<td>Naive-Bayes</td>
<td>+1.8%</td>
</tr>
<tr>
<td>GP-ensemble</td>
<td>+8.5%</td>
</tr>
</tbody>
</table>

4. REFERENCES


1 www.cs.umn.edu/~metis